Professor Mukadam

CS523 - Big Data Technology – Final Project

Part 1 :

Context: Covid19 data is ingested into system by Kafka, Spark Streaming read data from Kafka then store it in HBASE.

Prerequisites:

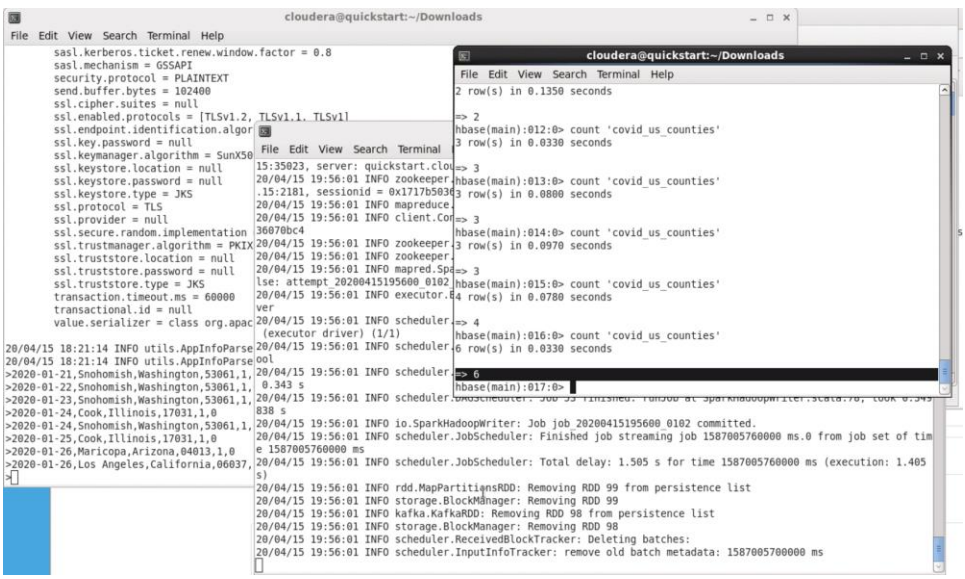
- Install Kafka as Part4
- Upgrade to Spark 2

Java Class: CovidDataFeed.java

Running Command: Running command: spark2-submit –master yarn –deploy-mode client –class CovidDataFeed finalproject-0.0.1-SNAPSHOT-jar-with-dependncies.jar

Shellscript: covid_streaming.sh

Screen shots:



Part 2:

Context: Covid19 data has been downloaded from https://github.com/nytimes/covid-19-data and imported to "covid19_us_counties" table on HBase. We will use Spark SQL to perform query data on HBase table.

Import data to Hbase: hadoop jar finalproject-0.0.1-SNAPSHOT-jar-with-dependncies.jar util.CovidHbaseTable

Java file: CovidSparkSQL.java

Running command: spark2-submit –master yarn –deploy-mode client –class CovidSparkSQL finalproject-0.0.1-SNAPSHOT-jar-with-dependncies.jar

ShellScript: covid_streaming_sql.sh

Screen shots:

```
20/04/14 09:41:24 INFO scheduler.DAGScheduler: Job 4 finished: show at Covid19Sql.java:29,
+-----------+----------+
|     county|sum(cases)|
+-----------+----------+
|  Snohomish|         5|
|     Orange|         1|
|Los Angeles|         1|
|       Cook|         2|
|   Maricopa|         1|
+-----------+----------+
```

Part 3:

Context: Covid19 data has been imported to "covid19_us_counties" table. We need to represent the data by visualize number of Covid19 cases in each county.

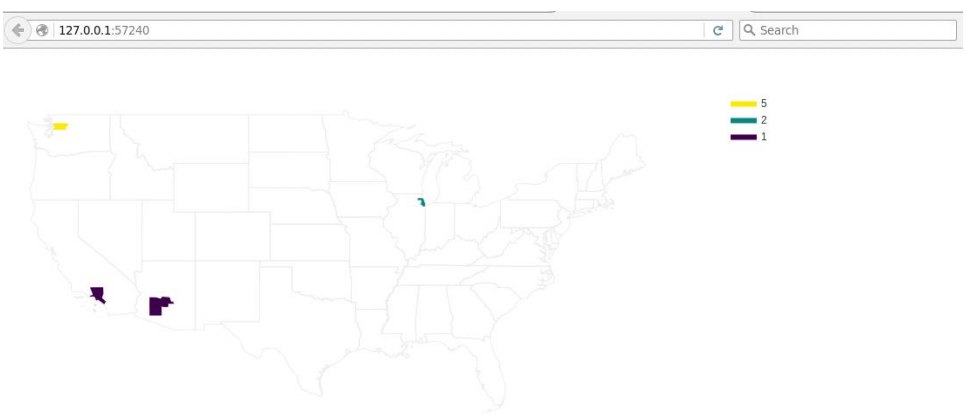
Prerequisites:

- Upgrade cloudera to support python 3
- Install happybase to conenct with HBase table
- Install plotly and plotly-geo to visualize the data

Python file: covidvisual.py

Shellscripts: covidvisualdata.sh

Screen shots:



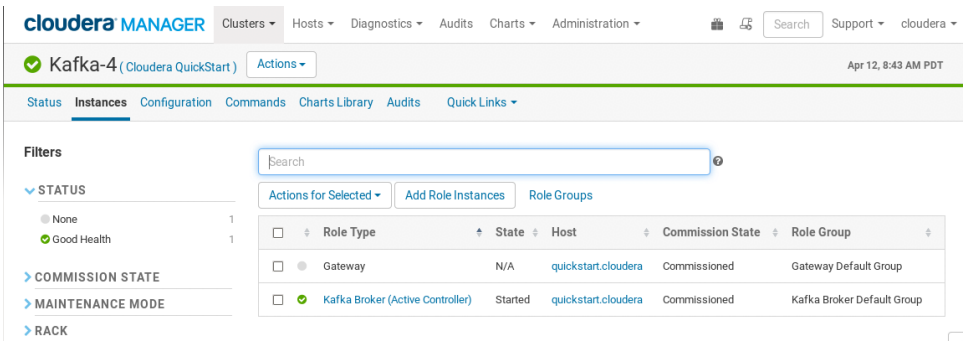
Part 4: Kaka Install and Integrate with other services

Install from parcels

Cloudera QuickStart

Parcel Name	Version	Status	Actions
ACCUMULO	1.7.2-5.5.0.ACCUMULO5.5.0.p0.8	Available Remotely	Download
	1.4.4-1.cdh4.5.0.p0.65	Available Remotely	Download
CDH 5	**5.13.0-1.cdh5.13.0.p0.29**	**Distributed, Activated**	Deactivate
KAFKA	4.1.0-1.4.1.0.p0.4	Available Remotely	Download
	3.1.1-1.3.1.1.p0.2	**Distributed, Activated**	Deactivate

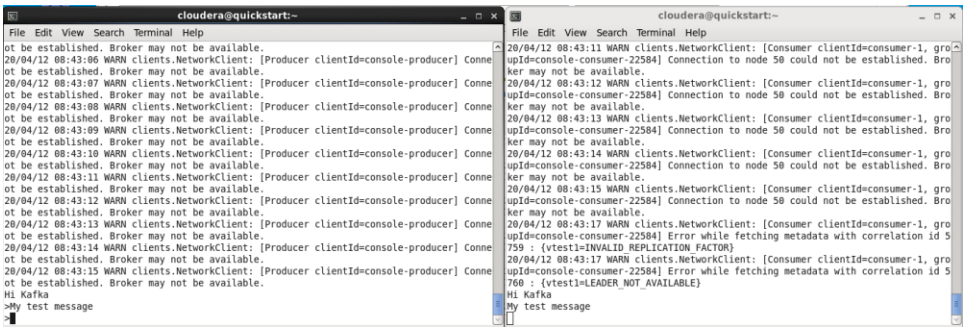
Service Running



Shellscripts:

- kafka_topic_create.sh
- kafka_producer.sh

Demo with Producer and Consumer



Data Source: https://github.com/nytimes/covid-19-data

Java Source code: finalproject.zip

Resources:

Upgrade to JDK8 for Cloudera

https://blog.clairvoyantsoft.com/upgrading-to-java8-on-the-cloudera-quickstart-vm-48855b1bc430

Install and configuration kafka

https://blog.clairvoyantsoft.com/installing-apache-kafka-on-clouderas-quickstart-vm-8245d8d0ebe5

Upgrade to spark 2

https://blog.clairvoyantsoft.com/installing-spark2-on-clouderas-quickstart-vm-bbf0db5fb3a9

Using Twitter api

https://towardsdatascience.com/how-to-capture-and-store-tweets-in-real-time-with-apache-spark-and-apache-kafka-e5ccd17afb32

https://www.edureka.co/blog/spark-streaming/

Databricks

https://community.cloud.databricks.com/?o=688775375094051#setting/clusters

Maven dependencies

https://docs.cloudera.com/documentation/enterprise/release-notes/topics/cdh_vd_cdh5_maven_repo_57x.html#concept_970_mcn_yk

Install Python 3

https://stackoverflow.com/questions/45803713/safely-have-two-versions-of-python-on-cloudera-virtual-machine-without-python-in/45804331#45804331

https://stackoverflow.com/questions/35246386/conda-command-not-found

```
conda create -y -n myproject 'python>3.6'

source activate myproject

python3
```